Understanding Memories of the Holocaust – A new approach to Neural Networks in the Digital Humanities

Introduction

This paper addresses an important challenge in artificial intelligence research in the humanities, which has impeded progress with supervised methods. The digital humanities do not have the resources to develop extensive test collections to train models, which has led to an overwhelming focus on unsupervised methods such as topic modelling or clustering (Jänicke et al., 2015). At the same time, computational modelling has become very popular in the humanities. This article introduces a novel method to creating test collections from smaller subsets. This method is based on what we will introduce as ‘distant supervision’ and will allow us to improve computational modelling in the digital humanities by including new methods of supervised learning. Furthermore, we present several computational methods to decipher the understanding our models generate about a text. This is especially important in our context, as our computational models are based on neural networks, which are among the most advanced approaches but are difficult to interpret. We finally present a number of insights of our use case covering Holocaust memories.

To demonstrate our new approach experimentally, we employ a real-life research question based on existing humanities collections. Sentiment analysis is the attempt to derive the emotion or 'sentiment' in texts. It has seen numerous applications in the digital humanities (Moreno-Ortiz, 2017, Lin, 2012). There are unsupervised and supervised versions of sentiment analysis. The unsupervised version is commonly used in the digital humanities (Moreno-Ortiz, 2017) and employs a dictionary of words to express sentiment. The supervised version of sentiment analysis learns from a training collection to model sentences structures, context, etc. expressing sentiments. Several highly-sophisticated applications are under development (Ravi and Ravi, 2015). However, the supervised model requires the availability of a domain-specific training collection. A commonly used collection is, for instance, the IMDB database of movies, which contains a short description of a movie and its review score. The supervised machine learning approach learns from the description what kind of score a new movie might achieve.

Sentiment analysis seems to be an excellent use case for our experiments. In terms of collections, we decided to focus on oral history collections, as they relate closely to sentiment analysis. Sentiment analysis should be useful to emerging digital oral history research to trace sentiments and memories in personal accounts of history. Oral history describes large and often informal collections recorded by individuals and groups that go beyond the official records of history by providing personal viewpoints and often emotions. With the emergence of social media, we are

widely expected to have a very large amount of oral history records related to future events. In this article, we do not cover such new records, but rely on traditional oral history interviews, where survivors of the Holocaust gave testimony about their experiences (Thompson, 2017).

In this paper, we present a methodology to combine supervised and unsupervised sentiment analysis in order to analyze the oral history interviews held by the United States Holocaust Memorial Museum. This methodology can help enhance available unsupervised methods with supervised techniques so that some of the above described limitations are overcome. There are many oral history collections of Holocaust testimonies, some of which like the Shoah Memorial collections belong to the largest digital cultural collections (Blanke, 2017). We will work with the textual transcripts of survivor interviews, conducted by the United States Holocaust Memorial Museum. These were given to us in the context of the European Holocaust Research Infrastructure (EHRI) project as a set of plain text files. EHRI was started in 2010 (Blanke and Kristel, 2013) and is currently funded under the Horizon2020 programme as a joint undertaking of Holocaust historians, archivists and specialists in the digital humanities. It offers access to Holocaust-related documents, often created under very difficult circumstances, helps preserve these documents and finally provides a range of digital methods to analyze them.

In total, the USHMM Holocaust testimonials consist of 1,882 text files of varying quality, going back to the 1980s. As oral records, they contain blank spaces, paragraphs, tabs, etc. and exhibit little organized structure, as one can expect, for instance, in official documents. Furthermore, the testimonies contain several unknown or misrepresented character encodings. Finally, the testimonies are all in English, but are still multilingual, as they are often not interviews with English-native speakers. It is common that the interviewees use their mother tongues to express, e.g., place or organization names. All these elements are typical features of many oral history collections, which make it difficult to computationally analyse them.

The testimonies are substantial in size. The longest testimony has 150,876 words and 14,611 sentences. The shortest testimony has 266 words and 34 sentences. The testimonies are, however, lexically not very diverse with an average diversity score of 8.79, measured by the average number of times a vocabulary term appears in a text. This is not untypical for transcripts of oral reports. This paper uses Python’s Pandas framework for data manipulation and integration (Raschka, 2015). Pandas allows for several different ways of quantifying texts and integrates with a number of machine learning and scientific tools.

Generating Test Data

The digital humanities lack larger annotated training collections that would allow it to apply supervised machine learning algorithms. This is not an uncommon issue, as most real-world applications do not have dedicated test collections. Recently a new approach called 'distant supervision' has been introduced, which addresses this issue (Mintz et al., 2009). Here, a classifier is learned given a smaller weakly labeled training set, which is often built automatically using intuitive labeling rules.[1] One of the commonly cited examples of distant supervision uses emoticons in Twitter data to create a training data set that is noisily labeled (Go et al., 2009). They extract emoticons in Tweets and use them as noisy sentiment labels. If the Tweet contains, for instance, a smiley, it will hint at a positive sentiment.

There are no emoticons in historical collections but we can use the unsupervised sentiment analysis method to derive initial sentiments. In this method, signal words summarized in dictionaries statistically represent a positive or negative polarity. Dictionary-based methods are a cost-effective way of determining sentiments and are thus very popular – provided there is a well curated sentiment dictionary. These lexica are, however, highly context-dependent. While dictionary-based sentiment analysis is therefore cheap in terms of computational efforts, its usefulness in practice is in doubt. Grimmer and Stewart (2013) list several known issues and conclude: 'Yes, dictionaries are able to produce measures that are claimed to be about tone or emotion, but the actual properties of these measures – and how they relate to the concepts their attempting to measure – are essentially a mystery.' (Grimmer and Stewart, 2013, 9).

The context-dependence of dictionary-based approaches weighs especially heavily on historical document or accounts, as sentiment lexica are generally based on modern word use and are often created out-of-context to analyse movie or financial sentiments. Rice and Zorn (2013) on the other hand have shown that the dictionary-based approach can have merits as a semi-supervised technique, where standardized sentiment dictionaries are developed into dedicated lexica for specific domains. We suggest a complementary approach contextualizing lexica by applying machine learning techniques that consider the whole body of knowledge in a collection as well as the specific language usages to express it. This way, we use the positive and negative signal words in sentiment lexica to derive training data.

Reliable sentiment analysis requires cost-intensive but also potentially highly subjective human-coding of many texts. As it is unrealistic to assume a labeled training and test dataset for Holocaust oral testimonies, our first step is to create a test collection using distant supervision. To this end, we first applied a dictionary-based sentiment analysis to develop a categorized corpus of those testimonies which contain the most negative sentiments and memories. As a dictionary, we employed a commonly used lexicon (Hu and Liu, 2004). We proceed by calculating the tf-idf weighting for all documents where we add up all the tf-idf values for positives terms contained in the documents and then subtract all tf-idf values for the negative terms. The final score is a positive value for largely positive memories or a negative value for largely negative ones.

[1] https://stats.stackexchange.com/questions/46685/distant-supervision-supervised-semi-supervised-or-both

We expect our collection to contain mainly negative memories, as they describe atrocities, war and forced migration. But the testimonies often exhibit more than one sentiment as they narrate the Holocaust experience as well as post-war stories. We therefore decided to first split the testimonies into paragraphs of 500 words each. We chose 500 as it is a good compromise over the length of the testimonies. Later on, we discuss how this choice could be improved in the future if we consider it as a hyperparameter. This way we created a larger corpus of over 46,000 text segments of testimonies. The alternative of using paragraph- or sentence-structures in the document is not feasible, because the documents do not follow a well-defined structure – as it common in oral records.

After applying our dictionary-based sentiment analysis to all testimony parts, it is not surprising that the collection of Holocaust testimonials is skewed towards negative sentiments, as Fig. 1 shows:

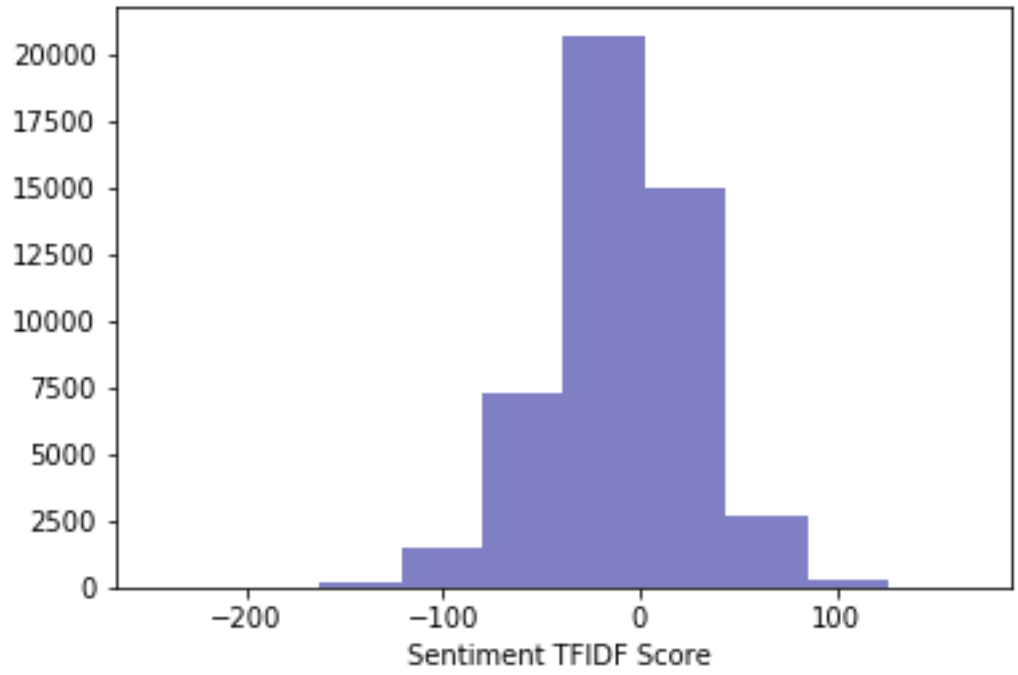

Fig. 1: Dictionary-based Sentiment Scores

To continue with our experiments, we need test and training samples, which are representative of positive and negative sentiments in the testimonies. We have chosen to only use the tail ends of the dictionary assessments and created a smaller test and training sample of the original corpus by selecting the 25% texts with the strongest negative sentiments and the 25% with the strongest positive sentiments. The resulting test and training corpus are 9,033 negative oral testimonies and 8,482 positive memories. The remaining 30,248 testimonies we ignore for the time being. This distribution has the additional advantage that positive and negative sentiments are evenly represented.

We then used the positive and negative sentiments to generate a new test collection for supervised learning using neural networks. Our aim was to create 8,000 additional sentiment documents for each classification. There are several ways of computationally creating such additional texts. We decided to use state-of-the-art Recurrent Neural Networks (Sutskever et al., 2011).

Using Recurrent Neural Networks to Generate new Training Data

Recurrent Neural Networks (RNNs) are popular to model sequence data such as time series and texts understood as sequences of words or characters (Karpathy, 2015). They develop a neural network model not just on current input but also previous ones. Recurrent nets use state-dependent information to perform tasks that other neural networks cannot. They look both forwards and backwards, with previous states as feedback loops to inform the decisions on the current states. Internally, they are modelled as a transition matrix not unlike the better-known Markov chains. In such a network, texts can be modelled as their sequences of characters or words. Given a series of characters, an RRN will use the first character to determine the most likely second character and so on. For instance, an initial q might lead it to infer that the next letter will be u, while an initial t might mean that the next letter will be h and so on.

RRNs can equally be used as generative as well as predictive models for text sequences. We will use them in predictive modelling later to predict the sentiment in all testimonies according to the model we generated. First, we employ the RRN to learn sequences of text and generate new likely sequences of text for the domain of testimonies. Please note, however, that calculating RRNs is a computationally very expensive process, which has limited our experiments. As said, we decided to generate 8,000 additional testimony fragments for each positive and negative sentiment class. On our standard Desktop hardware (dual-core with 16 GB RAM), even the generation of a model for 16,000 new testimonies took almost 10 hours. However, we think that the generative models have proven useful not only to understand the predictive model but the domain of oral testimonies themselves.

One epoch in neural network modelling is one forward and one backward pass of all the training examples. For a model trained with three epochs, we get relatively decipherable texts, though they still contain lots of character mistakes: ‘The so here in the talk the camp. This was way that’s want to allitions, the pratole alone, will, you diln my family to to the the fartion. I went them. I talk to sister of that the war and hease the tirans’. At the end of our training with 40 epochs, character mistakes are largely gone, and the text is not perfect but readable with logical chains of characters: ‘I have to be one and the partisans. I would say, I dont think that it was not to the cousin of a time, their sate they tell you, in there. And I didnt want to gress on it, I terribed the people was always were.’ With further training the readability could be further improved, but the generated text should be good enough for training positive and negative memories. In fact, in generating automated training collections, we want these kinds of errors, as they represent diverging texts.

After adding all the 16,000 new documents to the original training set, we ended up with over 33,000 documents to train a neural network to predict sentiments.

Predictive Modelling

We have chosen Recurrent Neural Networks (RNNs) again to provide the predictive model (Mesnil et al., 2013). In (Blanke, 2018), we discuss a new approach that we call ‘Predicting the Past’ to apply predictive techniques to past data and evaluate a range of models. We could rely on an extensive (non-)textual test collection and

focus on the interpretation. Here, we concentrate on experimenting with a new approach to developing test collections and would like to continue our discussion of RNNs as state-of-the-art in sequence modelling. An RNN will be used to model the language of the memory sentiment in the testimonies.

As a predictive algorithm, an RNN indexes each unique word in the testimonies with an integer number. Each sentence and document in the data is thus an array of integer indexes that represents sequences of words. The predictive RNN models these sequences rather than the character sequences the generative RNN targets. To ensure that the arrays are of the same length so-called padding is applied, which fills missing entries with 0s. Next to this padded indexing, we need two more standard pre-processing steps. Firstly, we use mini-batches, as it is common for neural networks. Secondly, as features we do not employ the indexed words directly but their embeddings (Bengio et al., 2003). Embeddings are vectors of real numbers that stand for the distributed representation of words. In our case, these numbers represent co-occurring words. Such deeper semantics improves the performance of the RRN.

We used a vanilla multilayer RRN consisting of several layers of neurons. The first layer is the word embedding layer, followed by a layer of long short-term memory (LSTM) cells. LSTM cells are the defining elements of RNNs. They remember inputs over arbitrary time intervals and function as the memory of RRNs. The LSTM layer is followed by layers of densely connected neural network cells. In our experiments, we used only one layer to avoid further training delays, as we struggled with training times on our standard hardware.

The preparation of the RRN is not the only pre-processing we have to do. The underlying testimonies are of very varying quality and heterogeneous, as they represent spoken everyday language. As described, words from different languages are often used interchangeably. The English word ‘camp’ appears together with the German word ‘Lager’. Furthermore, the transcripts contain several non-Unicode letters. Other errors include punctuation directly attached to words or incomplete words that the transcription could not identify. While we accepted non-English words to preserve the character of oral testimonies we removed obviously incomplete words. Furthermore, we applied further standard text pre-processing such as lower-casing and English stop word removal. We did not apply stemming, as in our experience it does not necessarily improve the text mining performance for historical documents.

In our experiment, we first ran the RNN model on our underlying non-enhanced dataset of 9,033 highly negative and 8,482 highly positive memories. It should be noted here that ‘highly positive’ memories should be considered with caution in this context. These are often normal memories, which appear positive only compared to the atrocities of the Holocaust. We will come back to this in the later discussions. The model contains several hyper-parameters such as the sequence length and the number of iterations for each training batch. For instance, with a sequence length of 200 or about half our document size and 50 iterations we reached a test accuracy of 0.73, while with a sequence length of 500 and the same number of iterations we reached 0.85 test accuracy. Both runs used 40 epochs and a 75-25% training and test data split. Unfortunately, training an RNN takes a long time on the standard hardware available to digital humanities. We had to limit the testing of different hyper-parameters and could not experiment with the structure of the RNN. In 10 different test runs, we achieved a test accuracy between 0.85 and

0.88. The best performing model (88% accuracy) in our ad-hoc tests used a sequence length of 500 with 250 iterations and an epoch size of 50.

We employed this model as a baseline to determine whether the enhanced, generated texts can improve the best model. Adding the generated documents, increases our dataset to 17,033 negative oral testimonies and 16,482 positive memories. We again apply a 75-25% training and test split of the data. With the baseline model, we reached an accuracy of 0.97 or 97%. This excellent performance is confirmed by the enhanced evaluation scores, as precision and recall were both 0.97, too. While this sounds like a great improvement, it still only corresponds to a percentage improvement of 8% using the enhanced dataset compared to the baseline model. However, a further improvement at the top of accuracy scores is difficult to achieve. We thus consider the experiment a success. The enhanced training set leads to a reliable model that we can use to decipher Holocaust memories.

The final step in our experiment is to apply the learned model to the whole original dataset of over 46,600 memories, as in (Blanke, 2018). In the digital humanities, we are not only interested in measuring algorithmic performance but also in using computational modelling to create new meanings a human reader might not see. We are interested to find the disagreement and misplacements the model identifies in the texts. In particular, we will evaluate whether the model might be better at determining the sentiment than the existing dictionary-based method. Next to the overall assignment of sentiment we also return the probability score.

The development of the RNN together with the embeddings improves the contextualization of the positive and negative sentiment words, the lack of which is often identified as the biggest disadvantage of the dictionary-based sentiment method (Grimmer and Stewart, 2013). A dictionary-based method might, for instance, assign a negative overall sentiment to the sentence: ‘We have overcome the losses of the war.’ ‘Loss’ and ‘war’ would both count for negative sentiment and outweigh the verb ‘overcome’, which might indicate a positive sentiment. This misclassification is based on the fact that only single words count for the classification, while the RNN and embedding should take the context of words and their typical order into account.

Let us compare next whether the RNN indeed improves the contextualization of sentiments by comparing its outputs with the standard dictionary-based sentiment analysis both qualitatively and quantitatively. We first compare sentiment readings by the dictionary and the RNN qualitatively. Afterwards, a full evaluation of a random subset of testimonies demonstrates that the RNN does perform significantly better than the dictionary-based method.

Quantifying and qualifying the RNN results

The challenges of the baseline dictionary-based method are best understood looking at an example how and why it misclassifies sentiments. A typical example is the second document part of RG-50.042.0003 (RG-50.042.0003_2)[2]:

> They [the Nazis] had advantage over other political parties such as the communists and democrats because they had a lot of money backing behind them and when they marched into a certain neighborhood, they had beautiful shiny boots and beautiful uniforms and they marched in military uh formation, whereas, democrats and the communists at that time, they were organized, but they were like a bunch of people running down the street, as compared to well-organized march of the Nazis.

The text is assigned a low negative count by the dictionary-based method. While it is a negative memory, words such as 'beautiful' and 'well-organised' mislead the classification. The neural network on the other hand assigns it a negative sentiment with a high degree of certainty, though the segment was not part of the original training dataset, as it was not in the most negative 25% of sentiments.

We can easily find more examples where the decisions of the dictionary-based method seem at least questionable. It is, for instance, not immediately clear, why RG-50.165.0129_trs_en.txt_7 achieves a strong negative score. It is a story of post-war emigration and the escape to Israel. The neural network, on the other hand, is confident that this is a positive experience. Another example for a negative dictionary-based assignment against a positive assignment by the RNN is RG-50.106.0113_trs_en.txt_5, which talks about a chaotic escape in September 1939: 'Oh, the next change I remember very, very distinctly was when the war broke out. […]. my father had a large car and my – I guess somebody else had a car, but I remember there was a lot of discussion about who was going to go in which car and how many things we can take and where we should go. And in the end we ended up in this car with my father's, I guess sister and my cousin who was much older, she was 14 […].' The dictionary-based method seems to confuse the chaos of the escape with an overall negative experience, while the neural network probably correctly assigns a positive sentiment. It aligns family experiences more strongly with positive memories, as we will later see.

Sometimes the dictionary-based classification is simply wrong. The following text part is assigned a very strong negative score:

> It was just a matter of circumstances that I survived after that. There was a uh, a very un, unexpected thing that had happened. While I was in in the childrens block, uh, I, […], they had a way of coming in and asking for volunteers. […]. Would you like to come. They were sending a group of children to a country where there is no war. That everything is going to be wonderful. Food is plentiful and everything is great. […]. (RG-50.042.0030_trs_en.txt_20).

One can see how this text part might be assigned a negative sentiment by the dictionary because of words like 'war' but it is a story of survival. Maybe worse, RG-50.549.05*0003 discusses Doctor Mengele's practices but is assigned a positive sentiment. Similarly, RG-50.042.0030_trs_en.txt_20 gets a positive score though it is a memory of the invasion of the Netherlands by the Germans and the removal of escape routes. On the other hand, a fairly harmless story about family relations (RG-

[2] RG-50.042.0003 indicates the finding aid identifier in https://collections.ushmm.org/oh_findingaids/. We use _2 to indicate the second document part of the reference document.

50.106.0113_trs_en.txt_5) is assigned a strong negative score by the dictionary-based method.

Comparing the RNN results to the dictionary-based ones next, overall over 24,000 negative experiences were found by the RNN compared to over 23,000 positive ones. For those documents that were part of the training collection, the neural network identified 9,024 negative sentiments compared to 8,466 positive ones. As a reminder, the dictionary-based method assigned 9,022 negative sentiments and 8,468 positive ones. This means that the two methods disagree only for 2 text components in the test collection. For instance, both methods agree that RG-50.106.0135_trs_en.txt_11 describes a negative experience: 'And the people -- again, they would starve them to death and then they would take people very weak, they call them musselmens, you know, just like skeletons, and they would make him go to Russian cemetery to bury them.' Similarly, RG-50.165.0044_trs_en.txt_4 is negative. It is about the experience of first escaping to France and then being arrested once the Germans had invaded.

Both the dictionary-based method and the neural network agree that RG-50.042.0003_trs_en.txt_16 is a positive experience. It is the description of the post-war return to Hamburg in Germany in 1972: 'We stayed there for I think 10 days, it was very nice, really very nice, the city of Hamburg is a beautiful city.' Many positive experiences in the testimonies are post-war. RG-50.106.0135_trs_en.txt_25 talks about a creative career in the US after the escape from Europe: 'Once [indecipherable] from United States run out, people from Canada came in. And I always was first fiddle, but I never could lead the band, you know, it always bothered me.' Both neural network and dictionary-based methods agree that the memory is positive, though at least some of the experiences seem ambivalent: 'And the hours are terrible. If they don't like you -- first of all, leave money, you can't -- if you have to go to the bathroom you can't do it, because it's money. And the managers play games with you, they don't like you.' But compared to the atrocities experienced during the war these bad work experiences are normal. The neural network has mainly learned to describe extreme emotions.

Of the texts that are not part of the training collection, the neural network assigned 15,168 positive sentiments and 15,042 negative ones. The neural network has learned to probably correctly assign a negative sentiment to the experience in RG-50.156.0042_trs_en.txt_12: 'To get on with the Gardelegen Flaming Death House, there were approximately about 1,000 various war prisoners that lost their lives here. They were mostly Russian, Poles, Jews, their own political radicals, and one American Negro soldier has been identified so far.' The dictionary-based method, on the other hand, believes that the following text has a positive sentiment: 'you left the ghetto and were sent to the concentration camp. A: Right. They separated the men and women.' (RG-50.106.0135_trs_en.txt_13). The neural network correctly assigns a negative sentiment. Both dictionary-based and neural network methods agree that the following family story is positive: 'in our house, I remember, we'd always have guests for a month or two, from [indecipherable] Friday night, my dad would always pick up a guest from si -- from synagogue.' (RG-50.106*0135_42).

This qualitative evaluation clearly indicates that the RRN performs better than the dictionary-based method. The improved performance is confirmed by our quantitative evaluation. To evaluate the performance of the neural network quantitatively, we randomly selected 500 testimony components and assigned positive or negative sentiment values to them. The manual assignments are in

agreement with 86.8% of the neural network ones. Precision is high with 94% but recall is worse with 80.6%. There are obviously sentiments which escape our model. The dictionary-based method faired a lot worse, if we were even able to assign a sentiment. For over 150 of the randomly selected document components the dictionary-based assignment was so ambivalent that we are not able to assign a sentiment value. If we included these, we were only able to achieve a 54% recognition of the actual sentiment using the dictionary. If we excluded all those the dictionary-based method could not decide upon, we are left with 328 texts only and an accuracy of 82%. The neural network method is a clear and significant improvement in either case.

Fig. 2 is the confusion matrix for the neural network method. The worst case is a negative prediction but a positive original sentiment. This is not surprising, as we deal with highly negative sentiments, and might even indicate a slight overfitting of negative sentiments.

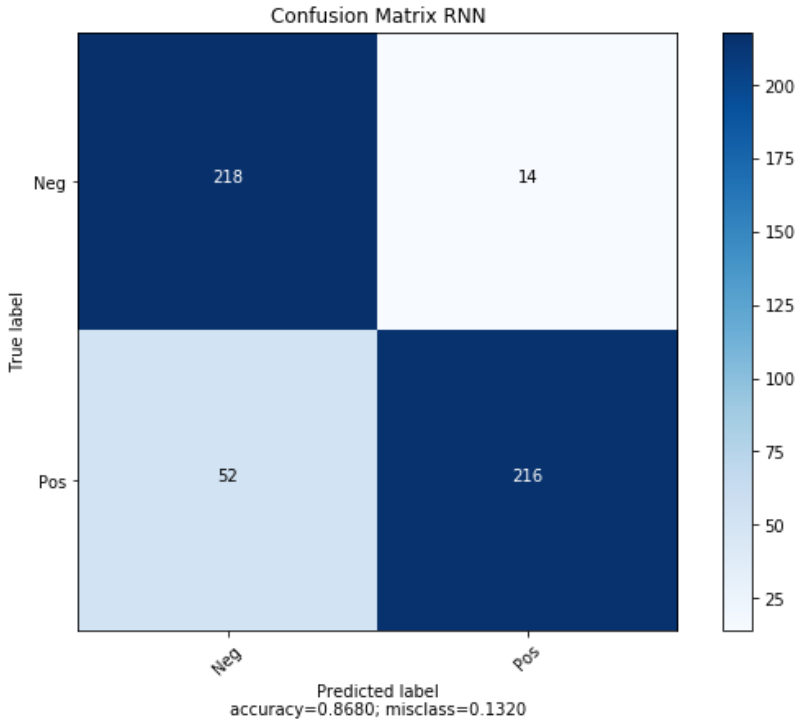

Fig. 2: Confusion Matrix of Neural Network

The neural network is closely trained on negative experiences. Based on our manual evaluation, RG-50.106.0065_trs_en.txt_1, for instance, counts for a positive experience. The text tells the normal experience of a Jewish child: 'Q: Did you have friends that were not Jewish? A: Some. Not too many. I lived in Jewish neighborhoods most of the time, but I did have some. Q: Did you experience any anti-Semitism as a child? A: Not really, because living in a Jewish neighborhood, I thought the whole world was Jewish.' Such normality has to count for positive experience within the context of the testimonies. RG-50.030.0669_trs_en.txt_19 is a typical example of a post-war occupation experience, which describes a difficult time. However, compared to the Nazi period it is normal and therefore 'positive'. RG-50.030.0467_trs_en.txt_70, on the other hand, is about the Israeli war with Jordan after the Second World War, which is linked to negative sentiments but again might

count for a normal life experience compared to Holocaust memories. RG-50.549.05.0007_trs_en.txt_22 discusses family relations in Terezín, which the dictionary-based method believes to be a positive experience. The neural network understands the negative context of a camp and assigns it a negative sentiment.

To understand these 'neutral' and therefore 'positive' experiences further, let us consider those text fragments, where the RNN assigned a sentiment probability of more than 0.4 and less than 0.6. These are the most uncertain sentiments. RG-50.233.0022_trs_en.txt_10, for instance, is a description of post-war life but fairly neutral: 'We lived in Chicago because Chicago _____ invited him for job. So we went to Chicago. Then to America, we were first in Italy and so to America.' RG-50.030.0424_trs_en.txt_16 is interesting, as the neural network disagrees with the dictionary-based analysis and marginally assigns it a positive sentiment with a probability of 0.56. It is a story about the war but also underground resistance: 'I had about 12 people in my underground, furriers. I was there, my brother was there with me. My brother was still with me at that time. A mother I didnt have, a father I didnt have, and what happened with the rest, I didnt know.' Again, the neural network identifies a family narration to be part of positive memories.

We have already discussed several examples, where the dictionary-based method arrives at the wrong conclusions. RG-50.549.05*0003, for instance, is about Mengele's experiments: 'And the Doctor, Doctor Mengele was a very tall fellow. And he touched one, two, and here I was laying in the middle. And he didn't touch me, because I was still warm.' The neural network is very confident that this text is about a negative experience, contradicting the dictionary-based score, which sees this as a positive memory. Similarly, RG-50.030*0056_3 is the story of Jewish life under the Nazis and the 'yellow star': 'you should always be identified and if somebody wanted to spit on you, kick you, shoot you, take you away, thats what you were to do.' The neural network believes this to be overall negative but with a low confidence score of 0.22 (with 0 standing for negative experiences and 1 for positive ones). This is probably a correct consideration, as the text is also about possible exemptions from wearing the yellow star. The dictionary-based method assigns this text a very high positive value – possibly because it is about playing as a child. But, this child experiences antisemitism: 'And she said, "No, she cant play with me tomorrow either." And I said, "Well, why not?" And she said, "Because youre Jewish."'.

However, the neural network is not perfect and gets several entries wrong. E.g., it is confident that RG-50.030.0289_trs_en.txt_25 is a positive memory, though the document is about the horrible experiences of a family in the Ghetto. While it is about individual family members, mothers, fathers and brothers and their experiences, their work life, etc., the memory is clearly not that of a normal family life: 'At that time, the Ghetto, nobody checked to see if somebody was ill or not ill or children because as long as you provided quite a number of workers, this labor department provided would say we need for tomorrow twelve hundred, fifteen hundred workers for this brigade.' The neural network has learned that family members and life appear frequently in the positive memories and was not able to link this particular example to the Ghetto. Similarly, the decision should be at least much more ambivalent on RG-50.042.0030_trs_en.txt_20, which is a story of survival but also about the horrific things that happened to people around the narrator: 'So tell me that story again. Did you see your mother at the same time? I saw my mother----------, but tell it to me as though you didnt tell me before. Well, uh, my, my father, like I say, my father came, he was standing in line, my mother was on the other side, on the other side of the trucks, (…).' The neural network again associates families

with positive experiences. Compared to earlier examples the context of atrocities is not strong enough to assign a negative memory.

Deciphering the RNN results

This section will investigate further which textual contexts are identified by the neural networks as components of positive and negative sentiments. We analyse the decisions made by the neural network using complementary approaches of computational semantics. We are interested in comparing our outputs to what can be considered to be the most advanced computational semantics in the digital humanities, which we use to decipher the neural network's decisions. One of the major criticisms of neural networks is that they make it difficult to understand why they arrive at conclusions. To understand negative and positive sentiment decisions better we therefore need to rely on different computational techniques. We employ three techniques to computationally summarize the negative and positive memories in the testimonies: Chi-Square of common words, word2vec and topic models. We are also encouraged by the insights we could gain on the structures of Holocaust memories. The next section demonstrates clearly that negative memories are semantically close to each other while the positive memories describe a whole range of experiences. This diffusion of positive memories will make it difficult for the neural network to describe Arendt's banalities of evil. The memories of survivors is dominated by extreme atrocities.

Chi-Square and Word2Vec

The 100 most common words for both negative and positive sentiments are very similar. They have in common that they are used to describe an experience. 'happened' is among the most common sentiment words, as are 'know', 'remember', 'years', etc. We need to therefore target those words that discriminate the two experiences rather than the 100 most common words and employ a Chi-Square metrics, an association score of words either in the class of positive or negative memories. Chi-Square is based on counting the number of times a word appears in the negative or positive testimonies compared to its total frequency in the whole collection as well as the total number of words in the most negative or positive testimonies compared to the overall size of the dictionary. This relation gives us the Chi-Square score of a word within the most negative or positive memories.

Using Chi-Square we develop a dictionary of English-language words that discriminate our two classes best. For both positive and negative memories, we calculated a list of unique words. Words in the list of negative memories include: 'nazi', 'targeted', 'septic', 'tot', 'tumbling', 'infect', 'decomposed', 'backbreaking', 'delaying', 'displeased', 'revulsion', 'taunting', 'compress', 'catastrophic', 'impersonal', 'scarcely', 'hysterics', 'chlorine', 'viciousness', 'ablaze', 'angrily', 'pained', 'harshly', 'plague', 'termin', 'hurted', 'gonorrhea', 'brute', 'executioner', 'carnage', 'worrisome', 'uneasiness', 'grievance', 'handgun', 'suffocate', 'abscess', 'repulsive', etc. Note that there were also words in the list that were not at first sight negative such as 'shopkeeper' or 'infinitely'. We removed these manually. In total, we ended up with 119 negative memory words. Words in the list of positive memories included: 'studious', 'handout', 'boldness', 'finesse', 'fortitude', 'lovable', 'symposium', 'ordination', 'rabbinate', 'mastery', 'sensational', 'valve', 'sociable', 'cholesterol',

'ligation', 'mosque', 'oasis', 'siesta', etc. Again, we curated the list and removed those without an obvious link to positive memories such as 'crossword' or 'kitchenette'. In total, there were 130 positive memory words.

With these highly discriminating memory words, some of the misclassifications we noticed earlier can be explained. RG-50.030.0289_trs_en.txt_25, for instance, is about the many forms of death in Ghettos but also about the (positive) experience of work: ‘Everybody was out working in the brigades. And they were going from house to house and anybody left who didnt go to work or was sick or bedridden or old or a child they took in the busses and took to the Ninth Fort. They took my mother then during that. When I came back from work, I found she was gone.’ RG-50.042.0030_trs_en.txt_20 on the other hand talks about survival in most difficult circumstances. Survival is closely linked to being saved according to our next word association investigation, which uses word embeddings.

In order to establish a better word association representation, we enhanced the list of most discriminative positive and negative words and employed again word embeddings. Our models, e.g., learn correctly that the series ‘man woman daughter house’ is not linked to the testimonies. We calculated two embeddings, one for each class of documents. For the negative memories, we are first interested in the related associations with ‘camp’. The model of negative memories includes the camps ‘Buchenwald’, ‘Auschwitz’, ‘Mauthausen’, ‘Dachau’, etc. It also delivers the German translation of ‘camp’, which is ‘Lager’, as well as related terms such as ‘prison’ and ‘compound’. Not related to ‘camp’ according to the model are ‘bless’ or 'educations' but the relationships are much weaker. Overall, it seems that the positive memories are much more difficult to describe for computational models. According to the embeddings of positive memories, ‘lovable’ is linked to generic terms such as 'easygoing' and 'good-natured'. In the positive memories, the concept ‘save’ is strongly linked to 'protect', 'raise', 'feed', 'defend', etc.

Fig. 3 shows the words strongly associated with negative memories and their relationships according to embeddings – mapped into a two-dimensional space using Principal Component Analysis (Raschka, 2015). The axis in Fig. 3 represent the two most significant principal components. Terms in similar contexts will thus appear close to each other. Each light-grey dot represents one term. Please note that there are many more dots than visible in Fig. 3, but we had to zoom in to the most relevant ones. The black dots are terms that are highly discriminatory for negative memories according to Chi-Square, the grey ones all the other terms. One can clearly see how most of the dots are closely concentrated in the centre around individual terrible experience in the Lager system. They are semantically close. Interestingly, the word Nazi seems to be outside this space, which might indicate that within the Lager system, Nazis were not seen as such anymore but as guards, murderers, etc.

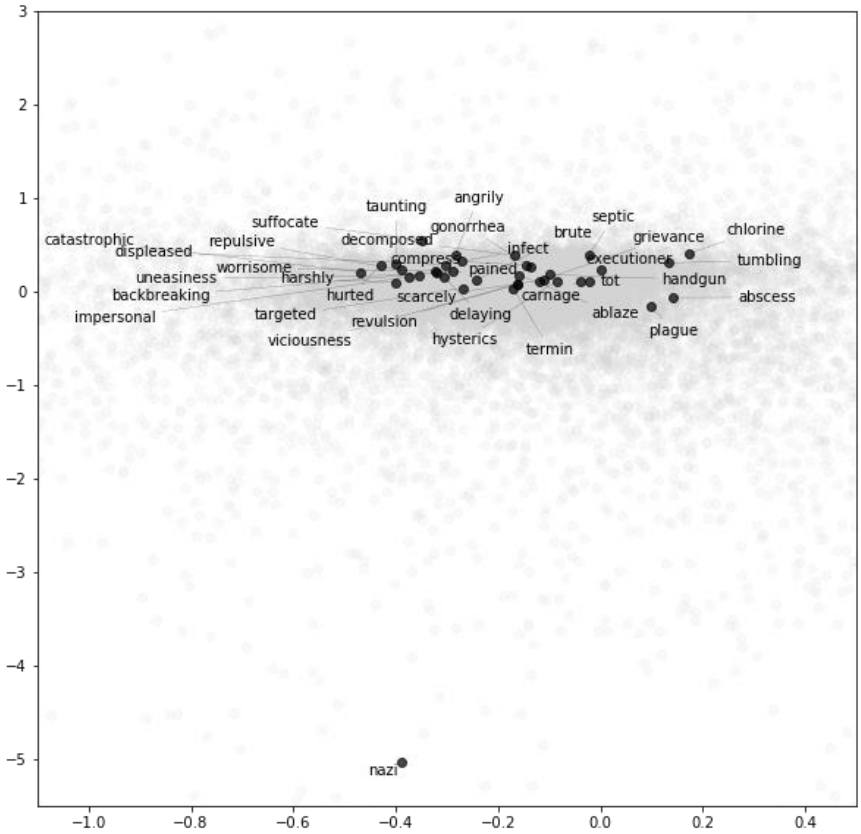

Fig. 3: Word2Vec of Negative Memories

Overall the grey dots and particularly the black ones are closely distributed in Fig. 3 around the camp centres. Fig. 4, on the other hand, shows the positive memories. The dots are much more wide-spread, which means that the positive memories are not so closely linked to specific experiences. We already suspected this earlier, as compared to the direct Holocaust memories all other memories can seem positive. Fig. 4 largely confirms that our methods target mainly extreme negative memories. The most discriminative terms describe very diverse feelings according to different parts of described lives. ‘Cholesterol’ is further away, as it is a generic life experience. It demonstrates how even discussions of common ailments can seem positive compared to the Holocaust experiences.

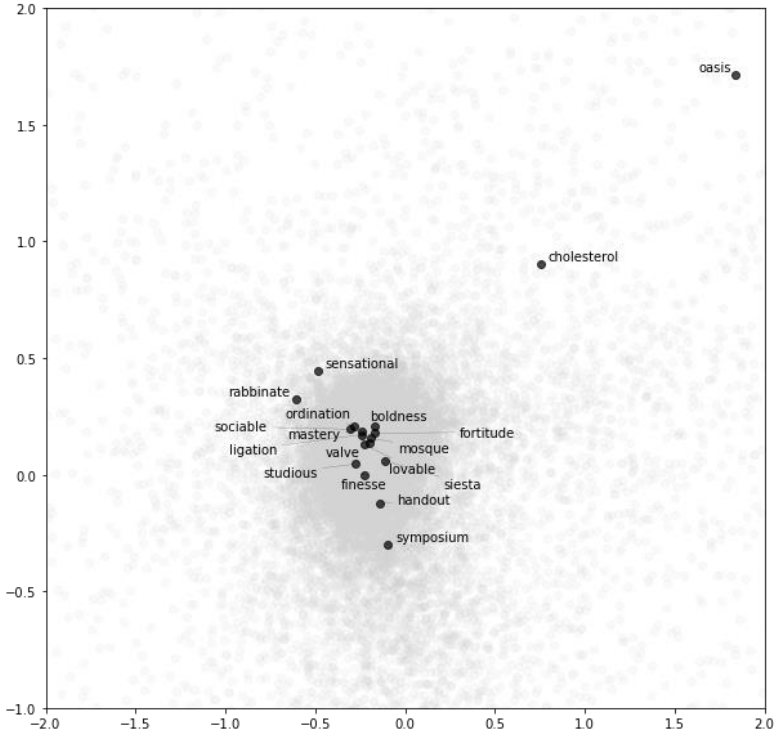

Fig. 4: Word2Vec of Positive Memories

The embeddings finally help to further understand the misclassifications of the neural networks. As described above, RG-50.030.0289_trs_en.txt_25 is not a positive memory as the RNN believes. It is about camp life but not directly about the concepts that discriminate the worst memories in the Ghetto such as ‘backbreaking’ or ‘handgun’. It describes forced labour experiences from an accounting point of view, which will have confused the RNN. Similarly, 50.042.0030_trs_en.txt_20 explains the line-up in a camp of family members, while the most discriminative negative memories are all linked to a state of degradation, disease and demise. They furthermore represent a very personal experience rather than the banality of evil that the accounting of workers in the Ghetto stands for, which does not care about workers as long as the numbers are correct. The banality of evil might have been how Hannah Arendt experienced Nazism after the war in the court case against of Eichmann (Arendt, 2006). It is, however, not the memory of the survivors according to our analysis. For them, the Holocaust is a very personal experience of demise.

The next section covers topic models, which again offer a further insight to how the memories are structured. The topic models make clear how family relations are strongly linked to positive memories even if they appear in the context of negative memories. To us, this demonstrates how personal memories in the testimonies are.

Topic Models

Embeddings are an improvement to Chi-Square, as they take into consideration the co-location context of terms in the testimonies. Topic models are also based on co-location and term association. They are very popular in the digital humanities as an

unsupervised learning technique to discover themes in a collection of documents (Mohr and Bogdanov, 2013). Topic models are statistical models of topics in a collection of documents. They follow the intuition that in a document about a topic, words are more likely to appear in the document frequently. In our documents, we would expect ‘camp’, ‘death’, ‘jewish’, etc. Furthermore, a document typically concerns multiple topics in different proportions. A topic model captures therefore a probabilistic distribution of topics in a document according to the overall collection it is part of. We use the Dirichlet allocation (LDA) algorithm, which has shown great results in practice (Blei et al., 2003).

Fig. 5 is the visualization of the distribution of 20 automatically generated topics in the collection using the t-distributed stochastic neighbor embedding (TSNE) dimensionality reduction (Maaten and Hinton, 2008). TSNE preserves spatial relationships between term vectors by keeping similar words close to each other while at the same time maximizing the distance between dissimilar terms. Fig. 5 demonstrates that topics visualized as colours are – generally speaking - combining similar terms.

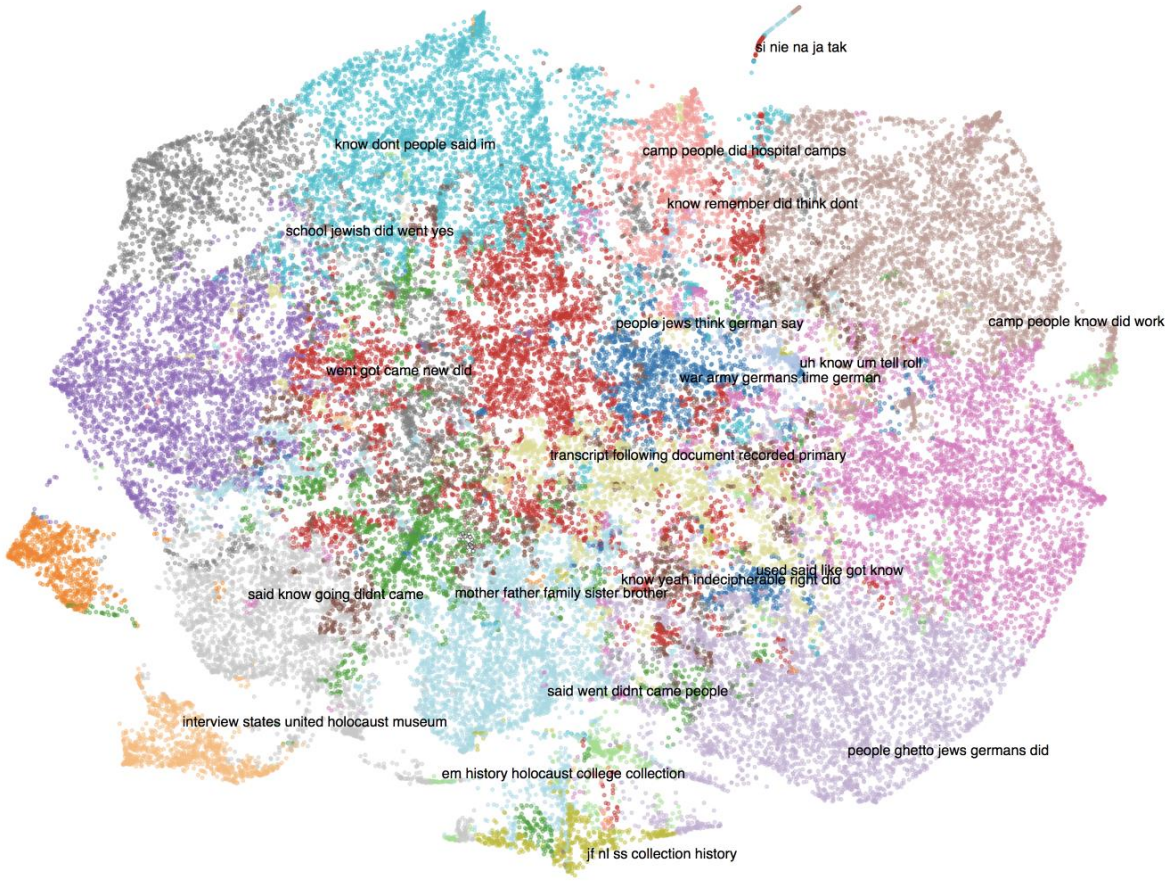

Fig. 5: 20 Topics in all Memories

The Figure shows the topic groups in 20 different colours and contains the 5 top words of each topic. While most topics are relatively distinct and occupy a particular section of the graph, others are clearly overlapping. For instance, we find parts of the archival memory (dark-brown) attached to all other memories. The memories in different languages can also be found attached to all other memories. Fig. 6 filters

the negative memories only. They are linked to all documents. While there are more strongly represented in some, they are still visible in all documents.

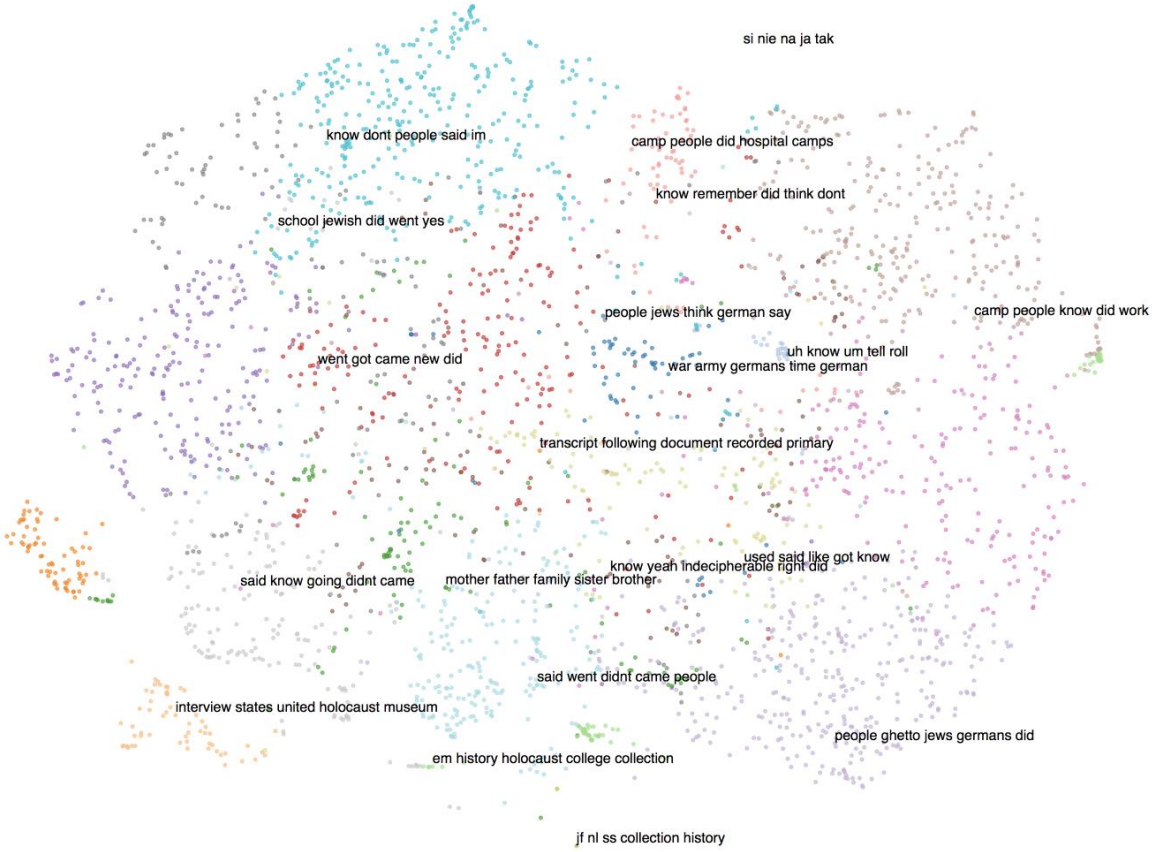

Fig. 6: 20 Topics in the Negative Memories

The topics describe memories we would expect from the testimonies. Topic 3 is probably about experiences of Jewish school children. Topic 6 is about German trains in the war, while Topic 8 portrays camps. Topic 9 summarizes relevant nationalities and Topic 10 is about family life. Topic 11 describes non-English words. Topic 16 is about Ghetto life, and Topic 17 about food and the lack thereof. We do not present the other 12 topics here, as they overlap with these 8 topics.

However, 20 is not the optimal topic number for the collection. To determine the optimal number of topics, we ran a final experiment using Konrad (2017). We chose a sample of 10,000 document components and ran a brute-force test with several topic modelling parameters. Based on three standard topic modelling evaluation measures assessing maximising likelihood and minimising KullbackLeibler divergence (Joyce, 2011), about 80 topics are optimal (Fig. 7). In the experiment, we developed a topic model with $\alpha = 50/k$, where k is the number of topics, and $\delta = 200/m$, where m is the total number of unique features (words) in the documents. In topic models, high α values mean documents belong to many topics. Higher δ values mean the topics contain many words.

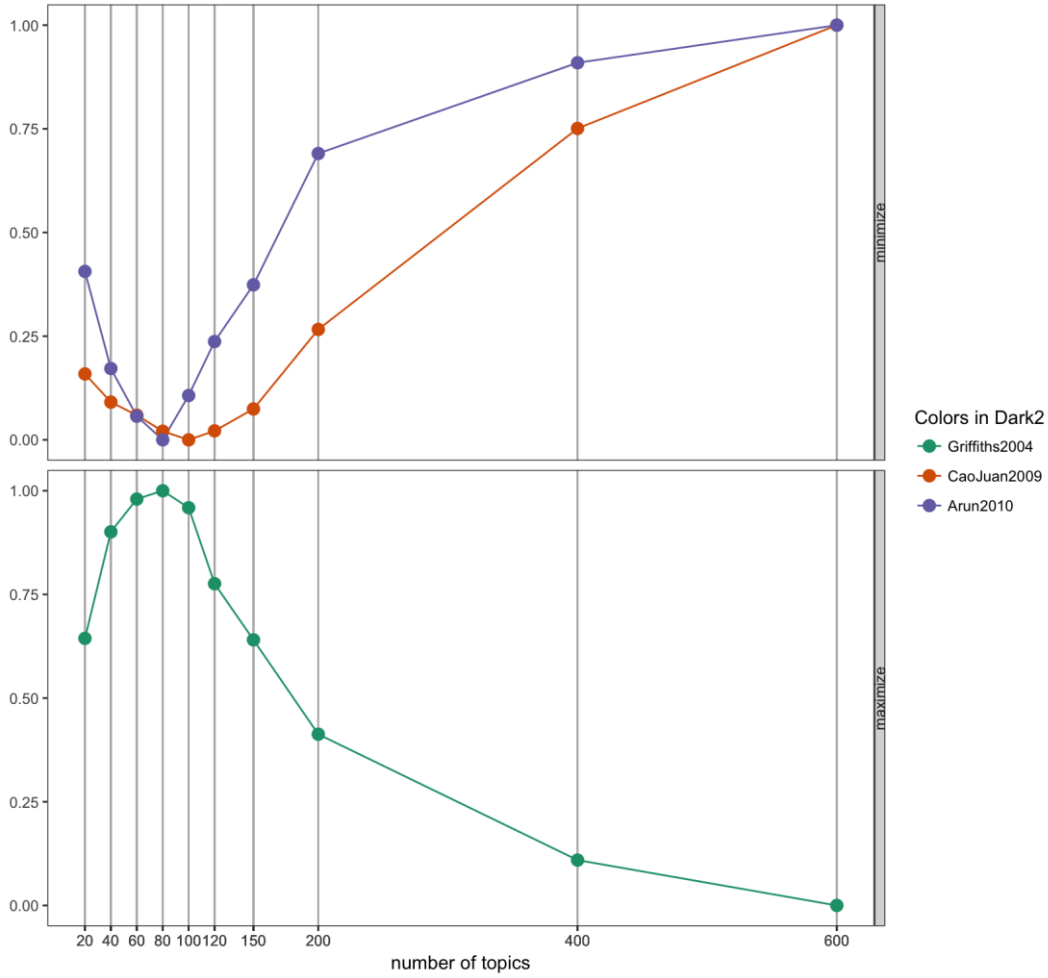

Fig. 7: Optimal Number of Topics

Unfortunately, 80 topics cannot be easily visualized anymore but we can use them to understand the RNN's decisions on the sentiment of memories better. To this end, we assigned the most relevant topic to each text and compared those that are assigned to texts with positive and negative sentiments. The most important topics in positive memories are Topic 40 [war time did going radio september january knew], Topic 46 [work factory working worked used like make people] and Topic 2 [roll ghetto page films kovno wentworth eh camera]. Negative memories were dominated by Topic 59 [polish poland warsaw know people jews jewish war], Topic 40 [war time did going radio september january knew] and Topic 63 [russian russians came war germans russia army people].

These topics explain some of the early discrepancies that the RNN and the dictionary produced. As described, RG-50.156.0042_trs_en.txt_12 is about the mass-murder of war prisoners, which lists many of the nationalities identified by the topic models to be part of negative memories because their countries and people were victims of the worst atrocities. The RNN believes this to be a negative memory but the dictionary-based method indicates that it is a positive one. The RNN thus agrees with the topic model. Furthermore, RG-50.030.0467_trs_en.txt_70 is about the war with Jordan after the Second World War and the reporting on it. The reporting of war rather than the war itself is according to Topic 40 and 46 linked to positive memories. Again, the RNN decision that this a positive memory is supported by the topic model.

As described above, RG-50.042.0030_trs_en.txt_20 is probably wrongly assigned a positive memory by the neural network because it discusses family relations. It is about the line-up of family members in a camp and the separation of children from their parents. According to the topic modelling, family memories described by keywords such as 'father, mother, family, parents, brother' are linked to both positive and negative memories. A seemingly positive memory is family Topic 75 [father family born mother town parents lived years], while a negative memory might be Topic 43 [father mother brother family sister parents died mothers]. Topic 75 is more prominent in positive memories with 182 compared to 43 counts in the negative memories. However, Topic 43 is also more prominent in the positive memories with 163 positive counts compared to 66 negative ones. This demonstrates to us that, as already suspected, family relations mainly contribute to positive memories even though this relationship might not be strong. Both Topic 43 and 75 are, however, not in the top 10 topics describing positive or negative memories, and neither are the other family-oriented topics: Topic 20 [children child mother little parents old kids years], Topic 25 [married husband years wife daughter son got met] and Topic 29 [sister letter got went brother didnt uncle aunt]. This weak association explains why in many examples above the neural network assigned a negative memory for texts about families, as the negative memory context was strong enough. Where the context was not strong enough, the family relations' higher count for positive memories than for negative ones misleads the RNN. Thus, the forced line-up of a family in RG-50.042.0030_trs_en.txt_20 is linked to a positive memory. Family relations seem to be indeed more prominent in positive memories though they are not exclusively found in them.

Conclusion

Artificial intelligence research is advancing fast in the humanities. However, a major current limitation is the lack of relevant training collections that allow the digital humanities to include more advanced methods into their practices. In researching memory experiences of Holocaust survivors, it is clearly unsatisfactory to rely on relatively brute-force dictionary-based methods. In this paper, we have demonstrated how digital humanities can take inspiration from related fields and use a combination of unsupervised and supervised techniques to develop a sophisticated contextualization of the language of Holocaust memories. Distant supervision has improved our computational models to a degree where we can seriously use them for an in-depth analysis of the language of Holocaust memories.

The article has introduced a new methodology, which starts from dictionary-based approaches to develop an initial evaluation of the Holocaust memories using unsupervised learning. Using Recurrent Neural Networks, we then proceeded to generate a larger training corpus of positive and negative memories. With this corpus, we were able to train a highly accurate model that qualitatively and quantitatively improved the baseline dictionary model. Based on the accuracy of the advanced model we are confident that we can analyse the Holocaust memories. To this end, we employed three advanced methods of computational semantics. These helped us decipher the decisions by the neural network and understand, for instance, the complex sentiments around family memories in the testimonies.

While we generally succeed with our objectives, several limitations remain mainly with regards to the testing and training infrastructures. Because the processing of neural networks requires advanced computational infrastructure, which we lacked, we were not able to apply appropriate testing of various hyper-parameters. These include neural network hyper-parameters such as sequence length or network architectures, which we should have tested further, but also the sentiment lexicon or the length of the testimony parts. For the lexicon, e.g., we relied on a single example, which might not be the best choice in our circumstances. Further experiments will improve our performance further. Overall, however, we have introduced a new approach for the digital humanities that will hopefully be useful for future research and open up supervised learning models.

Acknowledgement

This work was funded by the European Union (H2020) as part of the European Holocaust Research Infrastructure project (https://www.ehri-project.eu/). In particular, we are grateful to the United States Holocaust Memorial Museum and Michael Levy, Director of Digital Assets Management and Preservation, for providing us with a copy of the survivor testimonies.

References

ARENDT, H. 2006. *Eichmann in Jerusalem*, Penguin.

BENGIO, Y., DUCHARME, R., VINCENT, P. & JAUVIN, C. 2003. A neural probabilistic language model. *Journal of Machine Learning Research,* 3**,** 1137-1155.

BLANKE, T. 2017. Cultural Analytics. *Encyclopedia of Big Data***,** 1-5.

BLANKE, T. 2018. Predicting the Past. *Digital Humanities Quarterly,* 12.

BLANKE, T. & KRISTEL, C. 2013. Integrating Holocaust Research. *International Journal of Humanities and Arts Computing,* 7**,** 41-57.

BLEI, D. M., NG, A. Y. & JORDAN, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research,* 3**,** 993-1022.

GO, A., BHAYANI, R. & HUANG, L. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford,* 1.

GRIMMER, J. & STEWART, B. M. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis***,** 267-297.

HU, M. & LIU, B. Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004. ACM, 168-177.

JÄNICKE, S., FRANZINI, G., CHEEMA, M. F. & SCHEUERMANN, G. 2015. On close and distant reading in digital humanities: A survey and future challenges. *Proc. of EuroVis—STARs***,** 83-103.

JOYCE, J. M. 2011. Kullback-leibler divergence. *International Encyclopedia of Statistical Science.* Springer.

KARPATHY, A. 2015. The unreasonable effectiveness of recurrent neural networks. *Andrej Karpathy, Available: http://karpathy.github.io/2015/05/21/rnn-effectiveness/* [Accessed 13/7 2018].

KONRAD, M. 2017. *Topic Model Evaluation in Python with TMToolkit* [Online]. Available: https://datascience.blog.wzb.eu/2017/11/09/topic-modeling-evaluation-in-python-with-tmtoolkit/ [Accessed 13/7 2018].

LIN, Y.-W. 2012. Transdisciplinarity and digital humanities: Lessons learned from developing text-mining tools for textual analysis. *Understanding Digital Humanities.* Springer.

MAATEN, L. V. D. & HINTON, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research,* 9**,** 2579-2605.

MESNIL, G., HE, X., DENG, L. & BENGIO, Y. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. Interspeech, 2013. 3771-3775.

MINTZ, M., BILLS, S., SNOW, R. & JURAFSKY, D. Distant supervision for relation extraction without labeled data. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, 2009. Association for Computational Linguistics, 1003-1011.

MOHR, J. W. & BOGDANOV, P. 2013. Introduction—Topic models: What they are and why they matter. *Poetics,* 41**,** 545-569.

MORENO-ORTIZ, A. Lingmotif: Sentiment analysis for the digital humanities. Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017. 73-76.

RASCHKA, S. 2015. *Python Machine Learning*, Packt Publishing Ltd.

RAVI, K. & RAVI, V. 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems,* 89**,** 14-46.

RICE, D. R. & ZORN, C. 2013. Corpus-based dictionaries for sentiment analysis of specialized vocabularies. *Proceedings of NDATAD***,** 98-115.

SUTSKEVER, I., MARTENS, J. & HINTON, G. E. Generating text with recurrent neural networks. Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011. 1017-1024.

THOMPSON, P. 2017. *The voice of the past: Oral history*, Oxford university press.